View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Journal of Biomedical Informatics 51 (2014) 152–164

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Lessons learnt from the DDIExtraction-2013 Shared Task

Isabel Segura-Bedmar *, Paloma Martínez, María Herrero-Zazo

Dpto. de Informática, Universidad Carlos III de Madrid, Leganés 28911, Madrid, Spain

ARTICLE INFO

Article history:
Received 17 February 2014
Accepted 8 May 2014
Available online 21 May 2014

Keywords:
Information extraction
Relation extraction
Drug interaction

ABSTRACT

The DDIExtraction Shared Task 2013 is the second edition of the DDIExtraction Shared Task series, a community-wide effort to promote the implementation and comparative assessment of natural language processing (NLP) techniques in the field of the pharmacovigilance domain, in particular, to address the extraction of drug–drug interactions (DDI) from biomedical texts. This edition has been the first attempt to compare the performance of Information Extraction (IE) techniques specific for each of the basic steps of the DDI extraction pipeline. To attain this aim, two main tasks were proposed: the recognition and classification of pharmacological substances and the detection and classification of drug–drug interactions. DDIExtraction 2013 was held from January to June 2013 and attracted wide attention with a total of 14 teams (6 of the teams participated in the drug name recognition task, while 8 participated in the DDI extraction task) from 7 different countries. For the task of the recognition and classification of pharmacological names, the best system achieved an *F*1 of 71.5%, while, for the detection and classification of DDIs, the best result was an *F*1 of 65.1%. The results show advances in the state of the art and demonstrate that significant challenges remain to be resolved. This paper focuses on the second task (extraction of DDIs) and examines its main challenges, which have yet to be resolved.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Pharmacovigilance is formally defined by the WHO as "the science and activities related to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems" [1]. One of the major aims of pharmacovigilance is the early detection of adverse drug reactions (ADRs), which are unintended and harmful reactions to drugs. Several studies point out that the number of ADRs has increased significantly in recent years [2] and are responsible for about 5% of all hospital admissions [3,4]. More seriously, ADRs cause more than 300,000 deaths per year in the USA and Europe [5,6]. As a result, ADRs are a direct cause of the increase in health care costs [2]. Thus, the pharmacovigilance process is considered vital by pharmaceutical companies and drug agencies due to the high and growing incidence of drug safety incidents as well as their high associated costs.

Healthcare professionals are responsible for recognizing and reporting side effects by spontaneous post-marketing reporting systems. However several published drug safety issues have shown that the adverse effects of drugs may be detected too late, when millions of patients have already been exposed to them [7]. This fact poses a serious problem for patient safety giving rise to a growing interest in improving the early detection of ADRs. Drug–Drug Interactions (DDIs), which can be defined as alterations in the effects of a drug due to the recent use or simultaneously one or more other drugs, are an important subset of ADRs. Although there are different databases supporting healthcare professionals in the detection of DDIs (such as DrugBank [8]), the quality of these databases is very uneven and the consistency of their content is limited, so it is very difficult to assign a real clinical significance to each interaction [9,10]. On the other hand, these databases do not scale well to the large and growing number of pharmacovigilance literature in recent years [10]. In addition, a large amount of the most current and valuable information is unstructured, written in natural language and hidden in published articles, scientific journals, books and technical reports [11]. Thus, the large number of databases with information on DDIs and the deluge of published research have overwhelmed most healthcare professionals because it is not possible to remain up to date on everything published about DDIs.

Therefore, there is an increasing interest in facilitating automated access to information relevant on DDIs described in biomedical texts. Information Extraction (IE) techniques applied to pharmacovigilance literature can be of great benefit in the pharmaceutical industry allowing the identification and extraction of relevant information and providing an interesting way of reducing the

* Corresponding author. Fax: +34 916249129.
E-mail addresses: isegura@inf.uc3m.es (I. Segura-Bedmar), pmf@inf.uc3m.es (P. Martínez), mhzazo@pa.uc3m.es (M. Herrero-Zazo).

http://dx.doi.org/10.1016/j.jbi.2014.05.007
1532-0464/© 2014 Elsevier Inc. All rights reserved.

Table 1
Numbers of the annotated entities in the DDI corpus.

	DDI-DrugBank	DDI-MedLine	Total
DRUG	9901 (63%)	1745 (63%)	11,646 (63%)
BRAND	1824 (12%)	42 (1.5%)	1866 (10%)
GROUP	3901 (25%)	324 (12%)	4225 (23%)
DRUG_N	130 (1%)	635 (23%)	765 (4%)
TOTAL	15,756	2746	18,502

Table 2
Numbers of the annotated relationships in the DDI corpus.

	DDI-DrugBank	DDI-MedLine	Total
EFFECT	1855 (39.4%)	214 (65.4%)	2069 (41.1%)
MECHANISM	1539 (32.7%)	86 (26.3%)	1625 (32.3%)
ADVICE	1035 (22%)	15 (4.6%)	1050 (20.9%)
INT	272 (5.8%)	12 (3.7%)	284 (5.6%)
TOTAL	4701	327	5028

time spent by healthcare professionals and researchers on reviewing the literature.

With the support of collaborative events such as BioCreative [12–15], BioNLP [16–18], i2b2 [19,20], ShARe/CLEF eHealth [21] and SemEval-2014 Task 7 Analysis of Clinical Texts[1] shared tasks, there has been significant progress in IE techniques in the biological domain. However IE technology applied to pharmacovigilance still remains quite unexplored compared to biology.

The extraction of DDIs from biomedical texts has gained popularity and has seen significant advances recently with the organization of the DDIExtraction Shared Tasks in 2011 [22] and 2013 [23]. The main goal of these community challenges is to provide a common framework for the evaluation of information extraction techniques applied to the extraction of DDIs from biomedical texts. While the first event in 2011 only focused on the identification of all possible pairs of interacting drugs, the 2013 edition also included, in addition to DDI detection, the classification of each DDI. Furthermore, a supporting task, the recognition and classification of pharmacological substances, was proposed in 2013.

In the latest edition of DDIExtraction, a total of 14 teams submitted runs for at least one of the proposed subtasks (6 of the teams participated in the drug name recognition task, while 8 participated in the DDI extraction task). In the drug name recognition subtask, the top scoring team reached an *F*-score of 71.5%. In the relation extraction task, the best system achieved an *F*1 of 65.1%. This paper focuses on the second task (extraction of DDIs). The aim of this paper is twofold: to provide a detailed description and discussion on the 8 participating systems in the second task, the extraction of DDIs, and to discuss the remaining challenges revealed by the error analysis on these systems.

This paper proceeds as follows: Section 2 describes the corpus used in the shared task; in Section 3 we give a detailed discussion of the participating systems; Section 4 presents the results obtained by the participating systems; Section 5 describes the major sources of errors in these systems; Section 6 presents a study as to whether the results are significant statistically; in Section 7 we propose an ensemble system of combining the top three methods using majority and union voting strategies; and finally, we close with a discussion in Section 8 of possible future steps of the DDIExtraction Shared Task.

2. The DDI corpus

The major contribution of DDIExtraction has been to provide a benchmark corpus, the DDI corpus. The DDI corpus is a valuable gold-standard for those research groups interested in the recognition of pharmacological substances or those specifically working in the field of DDI relation extraction. It consists of 792 texts selected from the DrugBank database (DDI-DrugBank dataset) and other 233 Medline abstracts (DDI-MedLine dataset) on the subject of DDIs. The corpus was manually annotated with a total of 18,502 pharmacological substances and 5028 DDIs, including both pharmacokinetic (PK) as well as pharmacodynamic (PD) interactions. Four entity types were proposed to annotate pharmacological substances: drug, brand, group and drug_n. The drug type is used to annotate those human medicines known by a generic name, whereas those drugs described by a trade or brand name are annotated as brand entities. The use of either generic or brand names depends on the drug information source. Thus, while generic names are used in medical and pharmacological textbooks as well as scientific medical journals, brand names are used in drug product labels. The group type was used to annotate groups of drugs. This type was included because the descriptions of DDIs involving groups of drugs are very common in texts. The last entity type, drug_n, refers to those active substances not approved for human use, such as toxins or pesticides. This type was included because interactions between drugs and substances not approved for human use are frequently reported in Medline documents.

DDIs were annotated at the sentence level and, thus, interactions spanning over several sentences were not annotated. Four different types of DDI relationships are proposed: mechanism (this type is used to annotate DDIs that are described by their pharmacokinetic mechanism), effect (this type is used to annotate DDIs describing an effect or a pharmacodynamic mechanism), advice (this type is used when a recommendation or advice regarding a drug interaction is given) and int (this type is used when a DDI appears in the text without providing any additional information). Tables 1 and 2 show the numbers of the annotated entities and relationships in each corpus, respectively.

Fig. 1 shows some examples of annotated texts in the DDI corpus. This figure has been taken from the WBI corpora repository.[2] The DDI corpus was adapted to the Stav format by the WBI team in order to be visualized using Stav on-line visualization tool [24]. The first example (A), taken from the MedLineDDI dataset, describes a DDI of mechanism type between a drug (named using a synonym different from its most common generic name, *fomepizole*) that inhibits the metabolism of a substance not-approved to be used in humans (*1,3-difluoro-2-propranol*). The second example (B) is also a sentence taken from MedLine and describes the consequence of a DDI (effect type) between *estradiol* (a generic drug) and *endotoxin* (a drug-n) in an experiment performed in animals. The last example (C) is a paragraph from the DDI-DrugBank dataset. Its first sentence describes the consequence of the interaction (effect type) of a drug, denominated by its brand name (*Inapsine*), when is co-administered with five different groups of drugs. The third sentence in C shows a recommendation to avoid these DDIs (advice type).

Inter-annotator agreement (IAA) was measured in order to assess the consistency and quality of the corpus as well as the complexity of the annotation task. Tables 3 and 4 present the results for the agreement per type of entity and per type of relationship, respectively. Results were calculated in terms of the standard Kappa statistic [25]. The overall IAA results suggest that the DDI corpus has enough quality to be used for training and testing NLP techniques applied to the field of pharmacovigilance. A detailed description of the DDI corpus can be found in [26].

[1] http://alt.qcri.org/semeval2014/task7/.

[2] http://corpora.informatik.hu-berlin.de/.

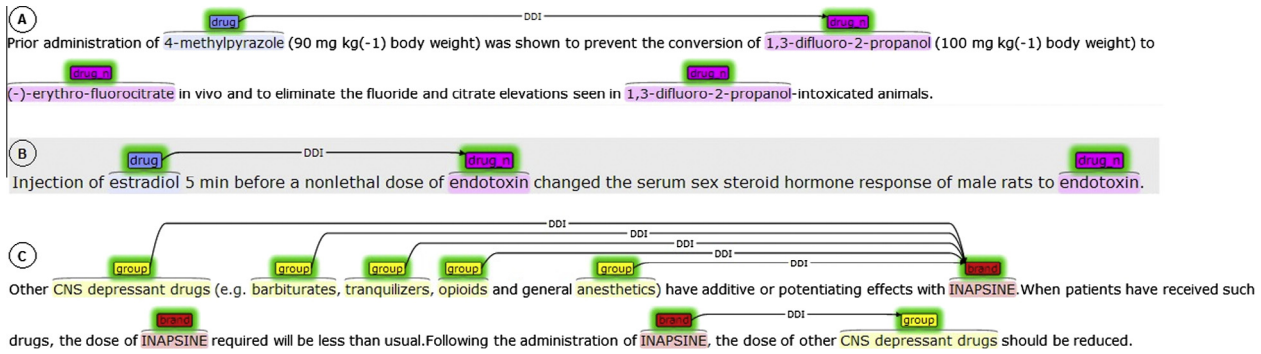

Fig. 1. Example sentences in the DDI corpus.

Table 3
IAA results of the annotated entities in the DDI corpus.

	DDI-DrugBank	DDI-MedLine
K_{DRUG}	0.9534	0.8467
K_{BRAND}	0.9569	0.8853
K_{GROUP}	0.9563	0.8299
K_{DRUG_N}	0.4422	0.8122
K	0.9104	0.7962

Table 4
IAA results of the annotated relationships in the DDI corpus.

	DDI-DrugBank	DDI-MedLine
K_{EFFECT}	0.7525	0.5548
$K_{MECHANISM}$	0.4214	0.5577
K_{ADVICE}	0.9428	0.5587
K_{INT}	0.9558	0.7252
K	0.8385	0.6213

3. Participating methods descriptions

This section reviews the 8 systems participating in the task of extracting DDIs and presents their results. For the evaluation of this task, the participants were given the test data with gold annotation only for pharmacological substances. The evaluation was then carried out by comparing the annotation predicted by each participant to the gold annotation. To simplify the task, the detection of DDIs was conducted at the sentence-level. The evaluation results are reported using the standard recall/precision/f-measure metrics, under different criteria: partial (only detection of DDIs) and exact (detection and classification of DDIs).

3.1. Fondazione Bruno Kessler team (FBK-irst)

The system consisted of two separate steps: first the DDIs were detected and second, the extracted DDIs were classified according to the proposed types (mechanism, effect, advice and int) in the guidelines task. In the DDI detection phase, filtering techniques based on the scope of negation cues and the semantic roles of the entities involved were proposed to rule out possible negative instances from the test dataset. In particular, a binary SVM classifier was trained using contextual and shallow linguistic features to find less informative sentences. A sentence is considered less informative when all its entities as well as its relation clues fall under the scope of a negation cue (no, n't, not). Less informative sentences were not considered in the relation extraction phase. Also, less informative negative instances were ruled out according to the following exclusion criteria: (1) if two mentions in a sentence refer to the same entity, this pair is not considered as a candidate DDI, (2) for any expression of the form "Drug1 (Drug2)", the pair was ruled out because both entities refer to the same entity (Drug2 is usually the abbreviation of Drug1), and (3) a candidate pair was ruled out when its two mentions had anti-positive governors with respect to the type of the relation. Anti-positive governors are words that tend to prevent mentions, which are directly dependent on those words, from participating in a certain relation of interest with any other mention in the same sentence [27]. We refer the reader to [28] for detailed description of anti-positive governors.

Once these negative instances were discarded from the test dataset, a hybrid kernel (combining a feature-based kernel, the shallow linguistic kernel (SL) [29] and the Path-enclosed Tree (PET) kernel [30]) was used to train a RE classifier. For the classification of the extracted DDIs, four separate models were trained for each DDI type (using ONE-vs-ALL). If none of the separate models is able to assign a class label to a predicted DDI, a default class label was chosen (for example, effect). The trained models were applied only on the extracted DDIs (by the DDI detection module) from the test dataset. Experiments on the training dataset showed that the filtering techniques improve both precision and recall with respect to applying only the hybrid kernel. This team achieved the best three submitted runs. The only difference between the three runs was the default class label which was "int", "effect" and "mechanism" for run 1, 2 and 3 respectively. The top run showed an $F1$ of 0.80 for DDI detection and 0.65 for DDI detection and classification.

3.2. Humboldt-Universität zu Berlin team (WBI)

The second best system was developed by the WBI team. The system relied on two step processes which first detected DDIs using ensembles of five different classifiers, and then the extracted DDIs were classified with one of the four proposed types. Several experiments were conducted combining the following eight machine learning methods: all-paths graph (APG) [31], the Shallow Linguistic Kernel (SL), SubTree (ST) kernel [32,33], Spectrum tree (Spt) [34], Turku Event Extraction System (TEES) [35], the case-based reasoning Moara system [36] and a self-developed feature based classifier (SLW), which is an extension of SL. Experiments were performed using 10-fold cross validation (CV) on the training set, and showed that the best results were achieved by the following majority voting ensembles: (1) Moara + SL + TEES, (2) APG + Moara + SL + SLw + TEES, and (3) SL + SLW + TEES. These ensembles were submitted as runs. This team was ranked second behind the FBK-irst team. Its best run was the third one, which yielded an $F1$ of 0.76 for DDI detection and 0.609 for DDI detection and classification.

3.3. University of Turku team (Uturku)

The third best team was the Uturku team. The TEES system was used to participate in both tasks: drug named entity recognition and extraction of DDIs. TEES is a machine learning system based on SVM, which was originally developed to extract events (and relations) in the BioNLP shared task. The event extraction is tackled as a graph generation task where nodes are keywords and edges are the words that connect nodes. The node detection task is similar to named entity recognition, while the edge detection task can be thought of as a relation extraction task. Deep syntactic features and information from external domain resources such as DrugBank or MetaMap [37] were used to model the Turku system. In run 1, the Uturku system was trained using only a feature set from syntactic parses. In run 2, DrugBank features were added to the syntactic features. Run 3 further extended run 2 with MetaMap information. The results of each run seem to be very close to each other. The best performance was provided by run 2 (an *F*1 of 0.696 for DDI detection and 0.594 for DDI detection and classification). While drug name recognition benefits from the use of domain knowledge resources, these external resources do not achieve a significant improvement in the relation extraction task. This may indicate that the extraction of DDIs seems to depend more on the syntactic interpretation of parse trees. TEES (version 2.1) is available for research purposes from http://bionlp.utu.fi/eventextractionsoftware.html. The authors also provided their DDI predictions for all DDIExtraction-2013 participants.

3.4. Complutense University of Madrid team (NIL_UCM)

The system was based on SVM using lexical, morphosyntactic and parse tree features. Information Gain ranker was used to eliminate the less informative features. The team submitted two different runs: (1) to train a SVM classifier with 5 categories (effect, mechanism, int, advice, null) and (2) to train a binary SVM classifer (DDI, non-DDI), and then, the extracted DDI were used to train a second SVM classifier with four categories (effect, mechanism, int, advice). The second run achieved better results (an *F*1 of 0.656 for DDI detection and 0.548 for DDI detection and classification) than the first one.

3.5. Carlos III University of Madrid team (UC3M)

The system was based on the SL Kernel. First, the SL was trained to distinguish positive instances from negative instances, and then, a SL model was trained for each DDI type. The SL kernel uses the following features: tokens, lemmas, PoS tags and entity types. In addition to the features listed above, the Anatomical Therapeutic Chemical (ATC) code of each drug name was obtained from the ATC system,[3] the drug classification system adopted by the World Health Organization (WHO). The team submitted two runs. In the first run, the team used the default setting of SL, while in the second one, the lemma feature was replaced by the ATC code of the drug. The first run achieved an *F*1 of 0.676 for DDI detection and 0.537 for DDI detection and classification. However, the use of ATC codes seems to give rise to a significant detriment to the performance with an *F*1 of 0.537 for DDI detection and only 0.294 for DDI detection and classification.

3.6. University of Wisconsin–Milwaukee and University of Pittsburgh team (UWM-TRIADS)

The system relied on two step processes: the first one detected DDIs using a binary weighted SVM classifier to discriminate positive instances (that is, DDIs) from negative instances, and then, a multi-class weighted SVM classifier was applied on the extracted DDIs (by the binary SVM) in order to classify each DDI. The team's hypothesis is that separating the detection and classification tasks into two different phases can help to handle the highly unbalanced class distribution. Texts were transformed into lower-case, drug names blinded and number were normalized. A feature set of lexical (such as bag of words and bigrams) and semantic features (synsets from WordNet [38]) was used to train a binary weighted classifier to discriminate positive instances from negative instances. Tokens were stemmed and lemmatized. The authors also used different stopwords lists of different size. The number of false positives was relatively high since the positive class was favored in the weighted SVM. The authors also defined a set of post-processing rules which were applied after the binary SVM classifier. For example, a rule consisted of discarding those pairs of interacting drugs referring to the same entity. Another example of a rule was when an interacting drug is a drug class of the other one; in this case, this pair should be ruled out since, in general, these pairs represent a hyponym/hyperonym relationship and not an interaction [39]. Other rules were aimed at detecting (without using any syntactic information) those pairs of drugs appearing in the same coordinative structure, since in general they are not interacting drugs.

Then, a multi-class SVM was trained on the set of extracted DDI classified by the previous binary classifier. In this case, the team proposed a rule that would assign the same type to all pairs obtained from drug mentions in a coordinative structure and other drug mention.

The team submitted three runs. The only difference between them was the size of the stopwords list used in each run. Experiments showed that the list of bigest size (263 stopwords) and the use of stems instead of lemmas achieve better results (*F*1 = 0.599 for DDI detection and *F*1 = 0.47 for DDI detection and classification) than the other settings.

3.7. Fraunhofer SCAI team (SCAI)

This system was based on the combination of three machine learning techniques: LibLINEAR [40] (linear SVM), Naïve Bayes and Voting Perceptron. While the first run was generated using only LibLINEAR, the second and third ones were based on majority and union ensemble learning strategies, respectively. All ML techniques used a rich feature vector consisting of lexical, syntactic and semantic features. The classification of extracted DDIs was performed by a post-processing step. This post-processing step uses a list of trigger words related for each type DDI which were manually created based on the observation of the MedLine dataset. The authors also applied an undersampling technique to balance the corpora and study its influence on the performance (only on the training dataset).

According to the official scores, their best result was obtained by run 3 (union voting strategy) with an *F*1 of 0.704 for DDI detection and 0.458 for DDI detection and classification. As regards the results for DDI classification, the system achieved the top score on MedLine (micro *F*1 = 0.42), however the system ranked at 5th position for DrugBank. This may be due to the trigger words collected were based on the observation of MedLine abstracts. Therefore, it would be advisable to define trigger words for each DDI type depending on the corpus.

3.8. University of Colorado team (UColorado_SOM)

This team used LIBSVM [41] trained with morphosyntactic, lexical and semantic features. The team applied the one-vs-all multi-class classification technique to handle the different DDI

[3] www.whocc.no/atc/,

Table 5
Machine learning techniques and tools used by the participating teams.

Team	ML technique	ML tool
FBK-irst	Feature-based kernel Path-enclosed Tree (PET) kernel Shallow Linguistic Kernel	 SVM-Light-TK toolkit jSRE
WBI-DDI	All-paths graph (APG) Shallow Linguistic Kernel TEES system SLW method	 jSRE Breeze Library
UTurku	SVM	SVM^{multiclass}
NIL_UCM	SVM	SMO Weka
UC3M	Shallow Linguistic Kernel	jSRE
SCAI	SVM	LibLINEAR
UColorado_SOM	SVM	LibSVM
UWM_TRIADS	SVM	LibSVM

types. Lexical and semantic features were used to train the classifiers. In run 2, the team also added features from TEES analysis provided by the UTurku team, and in run 3, features used in run 2 along with a list of interaction words were used as feature set. Their best run was the third one, achieving an *F*1 of 0.491 for DDI detection and 0.336 for DDI detection and classification.

3.9. Discussion

A common characteristic of all participating systems was the use of SVMs. While most systems used feature-based methods, only three teams (FBK-irst, WBI-DDI, UC3M) applied kernel-based methods which in general achieved better performance than the feature-based ones. Unlike feature-based methods, kernel-based methods do not require the explicit definition of feature vectors. A kernel-based method contains a kernel function and a kernel learner. A kernel function is a function that computes the similarity between two instances (for example, drug pairs). A kernel learner (such as SVM) is a learning algorithm which performs a learning task in a feature space. Table 5 gives a detailed view of all the ML techniques and tools used by the participating systems.

Most participating systems separate the learning problem into two stages: first the DDIs are detected and then they are classified into one of the types proposed in the guidelines. The only exceptions were the UTurku and NIL_UCM team. The TEES system, developed by the UTurku team, uses a multiclass SVM on a rich graph-based feature set. The NIL_UC3M team trained a multi-class SVM classifier with 5 classes (mechanism, effect, advice, int and null for negative instances). The NIL_UC3M also developed an approach in which the DDI detection and classification stages were separated. The evaluation on test dataset showed that the two-stage approach yielded better results than those achieved by the multi-class classifier.

As regards the two-stage approaches, the first stage, the detection of DDIs, is always performed by a binary classifier responsible for distinguishing between negative and positive DDIs instances. Most teams treated each DDI type as a single classification subproblem (one-vs-all). The SCAI team was the only one that did not use any machine learning techniques in the classification task. DDI instances detected in the previous step were classified according to a set of trigger words related with each type of DDIs.

As regards the natural language processing (NLP) tools often integrated into the participating systems, stemming, POS tagging and syntactic parsing were the most common ones. Stanford parser tools [42] were widely used by most systems. Around half of the participating systems used the Charniak–Johnson parser [43] with David McClosky's biomodel [44] trained on the GENIA corpus and unlabeled PubMed articles. From the results of the FBK-irst, WBI and UTurku teams, we can conclude that the parsers for the biomedical domain provided better performance than parsers trained for a general domain.

Table 6
NLP tools and other resources used by the participating teams.

Team	NLP tools	Knowledge resources
FBK-irst	Charniak–Johnson reranking parser McClosky's biomodel Standford parser BioEnEx (NER tool for diseases) [45]	
WBI-DDI	Charniak–Johnson reranking parser McClosky's biomodel Standford converter BioLemmatizer [46]	 DrugBank Phare Ontology [47]
UTurku	Charniak–Johnson reranking parser McClosky's biomodel Standford parser tools MetaMap	 WordNet DrugBank
NIL_UCM	TreeTagger [48] PaiceHusk Stemmer [49] Standford parser tools MetaMap	
UC3M	GATE Standford Parser plug-in	ATC system
SCAI	Porter Stemming algorithm [50] Charniak–Lease parser [51]	
UColorado_SOM	Genia Dependency Parser [52] Porter Stemmer	OpenDMAP [53] WordNet
UWM_TRIADS	Standford NLP tools Dragon tool [54] (a lemmatizer)	FDA Drug classification[a]

[a] http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162549.htm.

Some systems also used additional elements, such as lemmatization (WBI and UWM_TRIADS teams), semantic parsing provided by MetaMap (UTurku and NIL_UCM teams) or disease named entity recognition (team FBK-irst). Negation detection was only used by one team (FBK-irst). Surprisingly, only half of the participating systems used external lexical resources such as dictionaries or ontologies. Table 6 shows the NLP components and external resources used by the participating systems.

None of the participating systems made use of any additional training data collections to develop their systems, which implies that all systems relied only on the training dataset provided by the task organizers.

4. Evaluation results

This section summarizes the evaluation results and provides detailed analysis and discussion.

For the evaluation, the test dataset with gold annotation only for pharmacological substances was released to participants. Then, the evaluation was conducted by comparing the annotation predicted by each system to the gold annotation. The evaluation results are reported using the standard recall/precision/f-score metrics.

Table 7 shows the results of the DDI detection task. These results are not directly comparable with those reported in DDIExtraction 2011 due to the use of different training and test datasets in each edition. However, it should be noted that there has been a significant improvement in the detection of DDIs: almost all

Table 7
Results for DDI detection task on test dataset.

Team	Run	DrugBank				MedLine				Overall			
		Rank	*P*	*R*	*F*1	Rank	*P*	*R*	*F*1	Rank	*P*	*R*	*F*1
FBK-irst	1	1	0.816	0.838	0.827	1	0.558	0.505	0.53	1	0.794	0.806	0.8
	2	2	0.816	0.838	0.827	2	0.558	0.505	0.53	2	0.794	0.806	0.8
	3	3	0.816	0.838	0.827	3	0.558	0.505	0.53	3	0.794	0.806	0.8
WBI-DDI	1	6	0.857	0.686	0.762	8	0.63	0.358	0.456	6	0.841	0.654	0.736
	2	5	0.874	0.696	0.775	12	0.651	0.295	0.406	5	0.861	0.657	0.745
	3	4	0.814	0.755	0.783	4	0.625	0.421	0.503	4	0.801	0.722	0.759
UTurku	1	10	0.846	0.614	0.712	20	0.724	0.221	0.339	11	0.841	0.576	0.684
	2	9	0.861	0.624	0.724	19	0.778	0.221	0.344	8	0.858	0.585	0.696
	3	8	0.843	0.638	0.726	15	0.658	0.263	0.376	9	0.833	0.602	0.699
SCAI	1	11	0.836	0.619	0.711	7	0.688	0.347	0.462	10	0.826	0.592	0.69
	2	12	0.837	0.617	0.71	17	0.686	0.253	0.369	12	0.829	0.581	0.683
	3	7	0.796	0.681	0.734	6	0.431	0.526	0.474	7	0.748	0.666	0.704
UC3M	1	13	0.656	0.758	0.703	13	0.392	0.421	0.406	13	0.632	0.725	0.676
	2	19	0.415	0.814	0.549	10	0.313	0.642	0.421	19	0.404	0.798	0.537
NIL_UCM	1	16	0.615	0.615	0.615	22	0.419	0.137	0.206	16	0.608	0.569	0.588
	2	14	0.673	0.688	0.68	21	0.548	0.242	0.336	14	0.667	0.645	0.656
UWM_TRIADS	1	17	0.525	0.689	0.596	11	0.415	0.424	0.419	17	0.517	0.664	0.581
	2	15	0.573	0.665	0.616	9	0.427	0.446	0.436	15	0.561	0.644	0.599
	3	18	0.465	0.746	0.573	5	0.387	0.63	0.479	18	0.458	0.735	0.564
UColorado_SOM	1	22	0.387	0.739	0.508	16	0.256	0.663	0.37	21	0.37	0.731	0.492
	2	20	0.391	0.765	0.518	14	0.28	0.663	0.394	22	0.378	0.755	0.504
	3	21	0.422	0.646	0.511	18	0.253	0.6	0.356	23	0.398	0.641	0.491

Table 8
Results for DDI detection and classification task on test dataset.

Team	Run	DrugBank				MedLine				Overall			
		Rank	*P*	*R*	*F*1	Rank	*P*	*R*	*F*1	Rank	*P*	*R*	*F*1
FBK-irst	1	3	0.654	0.672	0.663	4	0.407	0.368	0.387	3	0.633	0.642	0.638
	2	1	0.667	0.686	0.676	2	0.419	0.379	0.398	1	0.646	0.656	0.651
	3	2	0.664	0.682	0.673	3	0.419	0.379	0.398	2	0.643	0.653	0.648
WBI-DDI	1	6	0.702	0.561	0.624	7	0.463	0.263	0.336	6	0.685	0.532	0.599
	2	5	0.707	0.563	0.627	12	0.488	0.221	0.304	5	0.695	0.53	0.601
	3	4	0.657	0.609	0.632	5	0.453	0.305	0.365	4	0.642	0.579	0.609
UTurku	1	9	0.723	0.525	0.608	18	0.517	0.158	0.242	9	0.714	0.489	0.581
	2	7	0.738	0.535	0.62	16	0.593	0.168	0.262	7	0.732	0.499	0.594
	3	8	0.706	0.534	0.608	13	0.5	0.2	0.286	8	0.694	0.502	0.582
NIL_UCM	1	12	0.54	0.541	0.54	22	0.387	0.126	0.19	12	0.535	0.501	0.517
	2	10	0.566	0.579	0.573	19	0.357	0.158	0.219	10	0.557	0.538	0.548
UC3M	1	11	0.518	0.598	0.555	15	0.265	0.284	0.274	11	0.495	0.568	0.529
	2	21	0.231	0.454	0.306	21	0.138	0.284	0.186	21	0.222	0.437	0.294
SCAI	1	15	0.546	0.404	0.464	1	0.625	0.316	0.42	14	0.551	0.395	0.46
	2	16	0.545	0.402	0.463	8	0.6	0.221	0.323	16	0.548	0.384	0.452
	3	14	0.513	0.439	0.473	6	0.31	0.379	0.341	15	0.486	0.433	0.458
UWM_TRIADS	1	17	0.407	0.534	0.462	10	0.309	0.315	0.312	17	0.4	0.513	0.449
	2	13	0.452	0.524	0.485	9	0.312	0.326	0.319	13	0.439	0.505	0.47
	3	18	0.361	0.579	0.445	11	0.247	0.402	0.306	18	0.35	0.562	0.432
UColorado_SOM	1	22	0.166	0.317	0.218	20	0.13	0.337	0.188	22	0.161	0.319	0.214
	2	20	0.258	0.503	0.341	14	0.196	0.463	0.275	20	0.25	0.499	0.334
	3	19	0.288	0.441	0.349	17	0.173	0.411	0.244	19	0.272	0.438	0.336

participants (except for the two worst teams) achieved an *F*-measure above 65.4% (the best *F*1 in DDIExtraction 2011). The increase in the size of the corpus made for DDIExtraction 2013, the inclusion of different types of documents and the quality of their annotations may have contributed significantly to this improvement. The best system (run 1 submitted by the FBK-irst team) had precision of 83.8% and recall of 83.8% (*F*1 82.7%) on DrugBank dataset, compared to its precision of 55.8% and 55.5% recall (*F*1 53%) on the MedLine dataset. It should be noted that there is almost a 30 point *F*-measure difference between the DrugBank dataset and the MedLine dataset. Indeed a common characteristic observed in all systems was the strong decrease in their results on the MedLine dataset compared to the DrugBank dataset. This may be justified by the different styles of the two sources. In the one hand, the texts taken from DrugBank are manually curated to provide brief descriptions of DDIs. Therefore, DrugBank contains short and concise sentences. On the other hand, the main topic of the scientific texts from MedLine would not necessarily be on DDIs. Moreover,

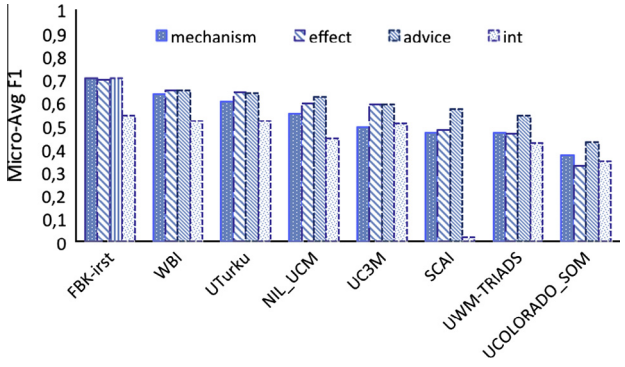

Fig. 2. Micro-Avg *F*1 scores by DDI type on the DrugBank test dataset.

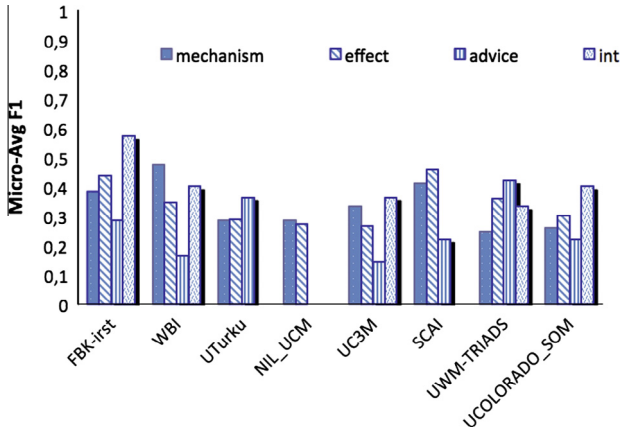

Fig. 3. Micro-Avg *F*1 scores by DDI type on the MedLine test dataset.

Table 9
Analysis of false negatives in the DrugBank dataset.

Error cause	FBK-irst	WBI	UTurku	Examples
Detection of coordinate structures required	14	23	17	E1, E2
Detection of appositions required	28	18	96	E3, E4
Unusual patterns for coadministration	11	18	27	E5, E6
Unusual patterns for DDI	28	51	20	E7, E8
Long DDI descriptions	6	27	20	E9
Unobvious DDIs	10	6	27	E11, E12
Resolution of percentages, dosages and temporal expressions required	8	4	26	E13, E14
Resolution of anaphora required	7	17	17	E15
Resolution of cataphora required	27	40	46	E16
Resolution of complex and compound sentences required	6	13	36	E17
Total	143	217	332	

Table 10
Example of false negatives in the DrugBank dataset.

ID	Example	DDIs not detected
E1	Several studies demonstrate a decrease in the bioavailability of **methyldopa**$_{e_1}$ when it is ingested with **ferrous sulfate**$_{e_2}$ or **ferrous gluconate**$_{e_3}$	(e_1, e_3)
E2	**Sulfoxone**$_{e_1}$ may increase the effects of **barbiturates**$_{e_2}$, **tolbutamide**$_{e_2}$, and **uricosurics**$_{e_4}$	$(e_1, e_2), (e_1, e_4)$
E3	Concurrent administration of **bacteriostatic antibiotics**$_{e_1}$ (e.g., **erythromycin**$_{e_2}$, **tetracycline**$_{e_3}$) may diminish the bactericidal effects of **penicillins**$_{e_4}$ by slowing the rate of bacterial growth	(e_3, e_4)
E4	Other inhibitors of the cytochrome P450 3A4 enzyme system, such as **antimycotic agents**$_{e_1}$ (e.g., **itraconazole**$_{e_2}$ and **miconazole**$_{e_3}$) or **macrolide antibiotics**$_{e_4}$ (e.g., **erythromycin**$_{e_5}$ and **clarithromycin**$_{e_6}$), may alter **oxybutynin**$_{e_7}$ mean pharmacokinetic parameters (i.e., Cmax and AUC)	$(e_1, e_7); (e_2, e_7); (e_3, e_7); (e_4, e_7); (e_5, e_7); (e_6, e_7)$
E5	The occurrence of stupor, muscular rigidity, severe agitation, and elevated temperature has been reported in some patients receiving the combination of **selegiline**$_{e_1}$ and **meperidine**$_{e_2}$	(e_1, e_2)
E6	The addition of **aspirin**$_{e_1}$ to **Streptokinase**$_{e_2}$ in the risk of minor bleeding	(e_1, e_2)

Table 11
Example of false negatives in the DrugBank dataset (cont. 2).

ID	Example	DDIs not detected
E7	There is usually complete cross-resistance between **PURINETHOL**$_{e_1}$ and **TABLOID**$_{e_2}$	(e_1, e_2)
E8	Concomitant treatment with **NEXAVAR**$_{e_1}$ resulted in a 21% increase in the AUC of **doxorubicin**$_{e_2}$	(e_1, e_2)
E9	Other drugs such as **cisapride**$_{e_1}$ or **pimozide**$_{e_2}$, which are metabolised by hepatic CYP3A isozymes have been associated with QT interval prolongation and/or cardiac arrhythmias (typically torsades de pointe) as a result of increase in their serum level subsequent to interaction with significant inhibitors of the isozyme, including some **macrolide antibacterials**$_{e_3}$	$(e_1, e_3), (e_2, e_3)$
E10	Certain **macrolides**$_{e_1}$ interact with **terfenadine**$_{e_2}$ and **astemizole**$_{e_3}$ leading to increased serum concentrations of the latter	$(e_1, e_2), (e_1, e_3)$
E11	**Furosemide**$_{e_1}$ and probably other **loopdiuretics**$_{e_2}$ given concomitantly with **metolazone**$_{e_3}$ can cause unusually large or prolonged losses of fluid and electrolytes	(e_2, e_3)
E12	Concomitant administration of **alcohol**$_{e_1}$ had a minimal effect on plasma levels of **mirtazapine**$_{e_2}$	(e_1, e_2)

these texts are characterized by a very scientific language and it is common the use of long and subordinated sentences. The error analysis (see Section 5) showed that the systems fall drastically for long and complex sentences. Another possible reason may be the different size between the two corpora. In addition, while the best system obtained balanced results in both precision and recall, the rest of the participants showed biased scores towards one or other metric.

As stated earlier, the use of biomedical parsers seems to provide better performance than parsers trained for a general domain, and the kernel-based systems in general overcame the feature-based ones.

The DDI classification task does not only consist of the identification of all possible pairs of interacting drugs, but also their classification. The results did not exceed an *F*1 of 65.1% (FBK-irst team) on the DrugBank dataset and 42% (SCAI team) on the MedLine dataset (see Table 8). These results clearly demonstrate that the identification of what type of information (such as an advice, an effect or information about the way the interaction occurs) is being used to describe a DDI may be a very complex task. As in the DDI detection task, all systems (except the runs submitted by the FBK-irst team) showed a marked disparity between precision and recall.

Figs. 2 and 3 show the results for each type of DDI on the DrugBank and MedLine test datasets, respectively. From each participant, we only select its best run. Fig. 2 suggests that some types

Table 12
Example of false negatives in the DrugBank dataset (cont. 3).

ID	Example	DDIs not detected
E13	Concomitant administration of **aspirin**$_{e_1}$ (1000 mg TID) to healthy volunteers tended to increase the AUC (10%) and Cmax (24%) of **meloxicam**$_{e_2}$	(e_1, e_2)
E14	All patients taking **NSAIDs**$_{e_1}$ should interrupt dosing for at least 5 days before, the day of, and 2 days following **ALIMTA**$_{e_2}$ administration	(e_1, e_2)
E15	Although **minoxidil**$_{e_1}$ does not itself cause orthostatic hypotension, its administration to patients already receiving **guanethidine**$_{e_2}$ can result in profound orthostatic effects	(e_1, e_2)
E16	**Drugs** which may potentiate the myeloproliferative effects of **Leukine**$_{e_1}$, such as **minoxidil**$_{e_2}$ lithium and **corticosteroids**$_{e_3}$, should be used with caution	$(e_1, e_2), (e_1, e_3)$
E17	**Mexitil**$_{e_1}$ does not alter serum **digoxin**$_{e_2}$ levels but **magnesium—aluminum hydroxide**$_{e_3}$, when used to treat gastrointestinal symptoms due to **Mexitil**$_{e_4}$, has been reported to lower serum **digoxin**$_{e_5}$ levels	(e_3, e_5)

Table 13
Analysis of false negatives in the MedLine dataset.

Error cause	FBK-irst	WBI	UTurku	Examples
Detection of coordinate structures required	3	0	7	E18
Detection of appositions required	5	6	3	
Unusual patterns for coadministration	17	17	15	E20
Unusual patterns for DDI	5	10	30	E21
Long DDI descriptions	3	4	2	E22
Unobvious DDIs	3	3	2	E23
Resolution of percentages, dosages and temporal expressions required	4	7	9	E24
Resolution of anaphora required	2	2	2	E25
Resolution of cataphora required	0	0	2	E26
Resolution of complex and compound sentences required	5	6	2	E27
Total	47	55	74	

of DDI are more difficult to classify than others on the DrugBank dataset, being the advice relationship being the easiest one. One possible explanation for this could be that recommendations or advice regarding a drug interaction are typically described by very similar text patterns such as '*DRUG should not be used in combination with DRUG*' or '*Caution should be observed when DRUG is administered with DRUG*'. The participating systems achieve very similar performance for the mechanism and effect relationships, while the int relationships seem to be the most difficult to extract. This may be because the proportion of instances of int relationship (5.6%) in the DDI corpus is much smaller than those of the rest of the relations (41.1% for *effect*, 32.3% for *mechanism* and 20.9% for *advice*).

5. Error analysis

The aim of this section is to perform a detailed error analysis with the objective of providing a road map for future work in the extraction of DDIs from texts. To this end, we focus on the study of the main source of errors produced by the systems developed by the following teams: FBK-irst, WBI-DDI and Uturku. For each team, we only analyzed their best runs. The reason for this choice is that these systems were the top-performing in DDIExtraction 2013.

Table 14
Example of false negatives in the MedLine dataset.

ID	Example	DDIs not detected
E18	**AAV2**$_{e_1}$-mediated retinal transduction is improved by co-injection of **heparinase III**$_{e_2}$ or **chondroitin ABC lyase**$_{e_3}$	(e_1, e_3)
E19	It is better to avoid prescribing isoenzyme CYP 2D6 inhibitors to women treated with **tamoxifen**$_{e_1}$ for breast cancer, especially **SSRI antidepressants**$_{e_2}$ such as **paroxetine**$_{e_3}$ and **fluoxetine**$_{e_3}$	$(e_1, e_3), (e_1, e_2), (e_1, e_3)$
E20	**Warfarin**$_{e_1}$ users who initiated **citalopram**$_{e_2}$, **fluoxetine**$_{e_3}$, **paroxetine**$_{e_4}$, **amitriptyline**$_{e_5}$, or **mirtazapine**$_{e_6}$ had an increased risk of hospitalization for gastrointestinal bleeding	$(e_1, e_2), (e_1, e_3), (e_1, e_4), (e_1, e_5), (e_1, e_6)$
E21	Reduction of PTH by **cinacalcet**$_{e_1}$ is associated with a decrease in **darbepoetin**$_{e_2}$ requirement	(e_1, e_2)
E22	In an in vitro assay, **lapatinib**$_{e_1}$ induced HER2 expression at the cell surface of HER2-positive breast cancer cell lines, leading to the enhancement of **Herceptin**$_{e_2}$-mediated ADCC	(e_1, e_2)

5.1. Analysis of false negatives

Tables 9 and 13 present the main causes for the false negatives in the DrugBank dataset and the MedLine dataset, respectively.

From Table 9, we can see that one of the most important factors contributing to false-negatives in DrugBank texts is the lack of cataphora resolution in the three systems. The resolution of the appositions in sentences, prior to the detection of DDIs, could allow to further improve the performance, particularly the FBK-irst and UTurku systems. Similarly, the resolution of anaphora and the detection of coordinate structures may also help to reduce false negatives, though fewer than the resolution of cataphoras and appositions. Another major cause of false negative is that many DDIs are described with very unusual text patterns. The high variability of natural language expression allows DDIs to be able to be composed using many different lexical and syntactic realizations. Classifiers have problems in detecting these cases since they are probably unrepresented in the training data. Tables 10–12 show some examples of false negatives in the DrugBank dataset.

Long, complex and compound sentences are other sources of false negatives. Many DDIs are described in long and complex sentences, which usually have a complex syntactic and lexical structure. Sentences with several embedded subordinated clauses are often encountered both in DrugBank and MedLine. Moreover, these sentences also pose a challenge to syntactic parsers due to their high levels of ambiguity. This may be one of the reasons why the methods using syntactic features from parsers (e.g. Standford parser) are not capable of dealing with these types of sentence. The FBK-irst system shows a lower rate of false negatives (only 2% are classified as long DDI descriptions and only 4% as complex and compound sentences) compared to the other two systems. In this case, the use of semantic roles, which were used to rule out

Table 15
Example of false negatives in the MedLine dataset (cont. 2).

ID	Example	DDIs not detected
E23	However, the evidence for a **calcium**$_{e_1}$ effect on **iron**$_{e_2}$ absorption mainly comes from studies that did not isolate the effect of calcium from that of other dietary components, because it was detected in single-meal studies	(e_1, e_2)
E24	Systemic and apparent oral **midazolam**$_{e_1}$ clearance were 24% (269 73 vs. 354 102 ml/min, $P = 0.022$) and 31%, respectively, lower in **cyclosporine**$_{e_2}$-treated patients ($n = 20$) than in matched tacrolimus-treated patients ($n = 20$)	(e_1, e_2)
E25	Acute administration of **hemantane**$_{e_1}$ or **doxycycline**$_{e_2}$ failed to influence locomotion in mice, while their combination normalized motor activity	(e_1, e_2)
E26	Regulatory agencies state that the combination of **clopidogrel**$_{e_1}$ and the CYP2C19 inhibitors **omeprazole**$_{e_2}$ and **esomeprazole**$_{e_3}$ should be avoided	$(e_1, e_2), (e_1, e_3)$
E27	Exposure to oral **S-ketamine**$_{e_1}$ is unaffected by **itraconazole**$_{e_2}$ but greatly increased by **ticlopidine**$_{e_3}$	(e_1, e_3)

Table 16
Analysis of false positives in the DrugBank dataset.

Error cause	FBK-irst	WBI	UTurku	Examples
Incorrect pair	57	60	36	E28
Annotation error	27	19	19	E29
Resolution of coordinated structures required	31	21	12	E30
Same drug	8	28	10	E31
Lack of evidence	41	21	9	E32
Resolution of apposition structures required	3	3	3	
Total	167	152	89	

possible negative instances, could be helping to overcome a wrong syntactic analysis.

Some sentences describe DDIs without giving an absolute certainty of their existence or using uncommon patterns. For example, in the sentence '*Lapatinib may have the potential to convert Herceptin-refractory to Herceptin-sensitive tumors in HER2-positive breast cancer by up-regulation of the cell surface expression of HER2*', it is even difficult for a human being to determine whether they are DDIs or not. The detection of dosages, numeric and temporal expressions can also help to improve the performance of the systems, since many sentences describe DDIs including additional information such as dosages, dosage regimen or percents of change of parameters, among others.

As regards the MedLine dataset (see Table 13), false negatives have similar error sources to those in the DrugBank dataset. The major cause of false negatives for all three systems is their inability to detect those DDIs described by patterns that are very scarce, even unrepresented, in the training data. This may be due mainly to the small size of the training dataset from MedLine. The detection of doses, numerical and temporal expressions also seem to be another significant problem which these systems have to face in order to improve their performance in the detection of DDIs. Anaphoras, cataphoras, coordinated structures and appositions have a much less significant effect on the false negatives in the MedLine dataset than in the DrugBank dataset. A possible reason for this could be that many texts in DrugBank provide descriptions of DDIs involving a drug and a list of drugs. The use of these linguistic structures is very common and useful in providing these kinds of description. Tables 14 and 15 show some examples of false negatives in the MedLine dataset.

5.2. Analysis of false positives

Tables 16 and 19 show the main causes of false positives in DrugBank and MedLine, respectively.

The major cause of false positives in DrugBank refers to sentences in which interacting drugs have more than one mention. The systems were able to detect that there was an interaction between two drugs, but failed to identify their mentions that were actually involved in this DDI. The first example in Table 18 (see E28) shows a sentence describing a DDI between *ibuprofen* and *ALIMTA*. We can see that both drugs appear twice in the sentence, but only their last two mentions are involved in the description of a DDI. However, the three systems failed to detect this DDI, because they proposed the pair formed by the first two mentions of the drugs. Annotation errors (see E29 in Table 17) are the second source of false positives in DrugBank. The candidate pairs were correctly detected by the systems, but were not annotated in the DDI corpus. Another cause of false positives was the systems' incapability to distinguish between drugs constituting a coordinate structure, and therefore, to recognize that they are not describing a DDI (see E30 in Table 17). Notably, one of the main sources of false positives would be fine with a simple rule that prevented mentions of drugs referring to the same drug which could be considered as a candidate DDI (see E31 in Table 18). The lack of evidence to confirm the existence of a DDI was another source of false positives (see E32 in Table 18). In fact, it was the main cause of false positives in MedLine (see Table 19 and Table 20).

6. Statistical analysis of significance

McNemar's significance test [55] is a χ^2-based significance test used to compare two groups, such as two classifiers or two

Table 17
Example of false positives in the DrugBank dataset.

ID	Example	FP	Gold DDIs
E28	Although **ibuprofen**$_{e_1}$ (400 mg qid) can be administered with **ALIMTA**$_{e_2}$ in patients with normal renal function (creatinine clearance 80 mL/min), caution should be used when administering **ibuprofen**$_{e_3}$ concurrently with **ALIMTA**$_{e_4}$ to patients with mild to moderate renal insufficiency (creatinine clearance from 45 to 79 mL/min)	(e_1, e_2)	(e_3, e_4)
E29	Careful monitoring of **phenytoin**$_{e_1}$ concentrations in patients receiving **DIFLUCAN**$_{e_2}$ and phenytoin is recommended	(e_1, e_2)	
E30	It may also interact with **thiazides**$_{e_1}$ (increased thrombocytopenia), **cyclosporine**$_{e_2}$ (increased nephrotoxicity), **sulfonylurea agents**$_{e_3}$ (increased hypoglycemic response), **warfarin**$_{e_4}$ (increased anticoagulant effect), **methotrexate**$_{e_5}$ (decreased renal excretion of **methotrexate**$_{e_6}$), **phenytoin**$_{e_7}$ (decreased hepatic clearance of **phenytoin**$_{e_8}$)	$(e_1, e_7), (e_2, e_7), (e_3, e_7), (e_4, e_7), (e_5, e_7), (e_6, e_7)$	

Table 18
Example of false positives in the DrugBank dataset (cont. 2).

ID	Example	FP	Gold DDIs
E31	Severe toxicity has also been reported in patients receiving the combination of **tricyclic antidepressants**$_{e_1}$ and **ELDEPRYL**$_{e_2}$ and **selective serotonin reuptake inhibitors**$_{e_3}$ and **ELDEPRYL**$_{e_4}$	(e_2, e_4)	$(e_1, e_2), (e_3, e_4)$
E32	There are no clinical data on the use of **MIVACRON**$_{e_1}$ with other **nondepolarizing neuromuscular blockingagents**$_{e_2}$	(e_1, e_2)	
E33	**Tetracycline**$_{e_1}$, a **bacteriostatic antibiotic**$_{e_2}$, may antagonize the bactericidal effect of **penicillin**$_{e_3}$ and concurrent use of these drugs should be avoided	(e_2, e_3)	(e_1, e_3)

Table 19
Analysis of false positives in the MedLine dataset.

Error cause	FBK-irst	WBI	UTurku	Examples
Incorrect pair	11	10	2	E34
Annotation Error	2	1	1	E35
Lack of evidence	35	13	3	E36
Total	48	24	6	

population samples. We applied the McNemar's significance test to compare the performance of the different runs and determine whether or not they differ significantly. Thus, for each pair of runs R_a and R_b, their corresponding models were performed on the test dataset, and the contingency matrix was then built for any pair of runs. The classification of each example in the test dataset by each model was recorded, counting the number of examples correctly classified by R_a and R_b (n_{11}), the number of examples correctly classified by R_a but not by R_b (n_{10}), the number of examples misclassified by R_a but not by R_a (n_{01}), and the number of examples misclassified by both R_a and R_b (n_{00}).

McNemar's test is based on a χ^2 goodness-of-fit test comparing the distribution of counts expected under the null hypothesis to the counts observed. The null hypothesis H_0 states that the two classifiers (runs) should have the same error rate (i.e., $n_{10} = n_{01}$). According to Dietterich [56], under the null hypothesis the following statistic (see Eq. (1)) is distributed as an χ^2 distribution with one degree of freedom.

$$\chi = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (1)$$

To test for significance, χ^2 was compared to the appropriate χ^2 table. Results with a probability greater than or equal to 0.05 are generally considered to be significant. Thus, the null hypothesis was correct if χ^2 was lower than $\chi^2_{1,0.05} = 3.841459$. In other cases, the null hypothesis could be rejected in favor of the hypothesis that the two runs produce different levels of performance.

First, we analyzed whether the runs submitted by a same team were statistically significant from each other. The runs submitted by the top four teams (FBK-irst, WBI, UTurku and SCAI) did not show statistically significant differences between the other runs of the same team. However, when comparing the two runs submitted by the NIL_UCM, a statistically significant difference did exist between their results. The first run used a multiclassifier with 5 categories (effect, mechanism, int, advice, null), while its second run followed a two-stage approach in which DDIs were initially detected using a binary classification, and then these detected DDIs were used to train a second SVM classifier with four categories (effect, mechanism, int, advice). Similarly, we also observed a strong statistically significant difference between the two runs submitted by the UC3M team. The second run used the ATC code instead of the lemma feature in the shallow linguistic kernel. This is the only difference between the two runs and it seems to give rise to a strong decrease in the performance. The only differences between the runs submitted by the UWM_TRIADS team were the size of the stopwords lists used in each run and the use of stems instead of lemmas. These changes led to statistically significant differences between the results of the three runs. As regards the runs submitted by the UColorado_SOM team, the use of a list of interacting words as an additional feature in the third run yielded the best results and also brought about statistically significant differences with the other two runs.

To study whether the runs submitted by the teams obtained statistically significant results with respect to the other participating teams, we only chose the best run from each team. Table 21 summarizes the χ statistic values for the pairwise comparison of the eight runs using the McNemar significance test (i.e., a total of 28 comparisons). Each cell of this pairwise comparison matrix represents the χ statistic value for a given pair of runs. Table 21 shows that in general most systems significantly differ from all others. The UTurku system does not significantly differ from the SCAI system. Similarly, the differences in performance between the UC3M system and the NIL-UCM system are not significant.

7. Ensemble

In order to investigate whether it is possible to improve the best scores obtained in DDIExtraction 2013, we built different ensemble systems by combining the submitted runs.

We chose to use a majority voting-based ensemble strategy due to its simple implementation. In this strategy, each system votes for a particular prediction (DDI or non-DDI), and the class with the most votes is selected as the final decision. In case of a tie, the final prediction was set to DDI, improving the recall (though at the expense of reducing the precision). This decision is based on the fact that all the participating systems have achieved better precision than recall. Therefore, in order to improve the *F*-measure, it is necessary to improve the recall. The main shortcoming of majority voting is that this strategy does not take in account that sometimes the minority predictions are correct.

Table 20
Example of false positives in the MedLine dataset.

ID	Example	FP	Gold DDIs
E34	**Moxifloxacin**$_{e_0}$ and **Lomefloxacin**$_{e_1}$ reacts faster with **sucralfate**$_{e_2}$ and **gelusil**$_{e_3}$ in acidic media whereas with **erythromycin**$_{e_4}$ in basic media and multi-minerals in neutral media	(e_2, e_3)	$(e_0, e_2), (e_0, e_3), (e_0, e_4), (e_1, e_2), (e_1, e_3), (e_1, e_4)$
E35	Improved parathyroid hormone control by **cinacalcet**$_{e_0}$ is associated with reduction in **darbepoetin**$_{e_1}$ requirement in patients with end-stage renal disease	(e_0, e_1)	
E36	On day 8, a single **panobinostat**$_{e_0}$ dose was co-administered with **ketoconazole**$_{e_0}$	(e_0, e_1)	

Table 21
χ statistic values using McNemar's test. The equivalent *p*-values are shown in parenthesis.

	WBI-3	UTurku-2	SCAI-1	UC3M-1	NIL_UCM-1	UWM-2	UCOLORADO-3
FBK-2	8.2 (0.00418904)	24.1 (0.00000091)	30.7 (0.00000003)	163.9 (0.0)	112.4 (0.0)	266.6 (0.0)	626.2 (0.0)
WBI-3	–	6.9 (0.00861957)	11.3 (0.00077507)	122.8 (0.0)	77.2 (0.0)	199.6 (0.0)	551.9 (0.0)
UTURKU-2	–	–	0.7848* (0.37567714)	47.7 (0.0)	41.6 (0.0)	143.5 (0.0)	496.3 (0.0)
SCAI-1	–	–	–	36.7 (0.0)	31.2 (0.00000002)	122.7 (0.0)	454.0 (0.0)
UC3M-1	–	–	–	–	0.42* (0.51693704)	28.8 (0.00000008)	270.3 (0.0)
NIL_UCM-1	–	–	–	–	–	40.8 (0.0)	293.7 (0.0)
UWM-2	–	–	–	–	–	–	132.7 (0.0)

* Not statistically significant differences.

Table 22
Ensemble systems results. 3bestruns: the three runs of FBK-irst; 6bestruns: the three runs of FBK-irst and the three runs of WBI; 9BestRuns: the three runs of the three top teams: FBK-irst, WBI and UTurku; 3BestRuns3BestTeams: the best runs of the three best teams: FBK-irst, WBI and UTurku; ExceptUColorado_SOM: best runs of each team, except UColorado; ExceptUWM_UColorado_SOM: best runs of each team, except UWM and UColorado; ExceptUC3M_UColorado_SOM: best runs of each team, except UColorado and UC3M; ExceptSCAI_UColorado_SOM: best runs of each team, except UColorado and SCAI; ExceptNIL_UCM__UColorado_SOM: best runs of each team, except UCM-NIL and UColorado.

Row	Team	DrugBank			MedLine			Overall		
		P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
1	Best system (FBK run 1)	0.816	0.838	0.827	0.558	0.505	0.53	0.794	0.806	0.8
2	All runs	0.8617	0.7681	0.8122	0.6923	0.3789	0.4898	0.8512	0.7303	0.7861
3	3bestruns	0.8161	0.8382	0.827	0.5581	0.5053	0.5304	0.7938	0.8059	0.7998
4	6bestruns	0.8004	0.8575	0.828	0.5618	0.5263	0.5435	0.7799	0.8253	0.802
5	9BestRuns	0.885	0.7489	0.8113	0.6596	0.3263	0.4366	0.8717	0.7079	0.7813
6	Best runs of each team	0.8485	0.7919	0.8192	0.678	0.4211	0.5195	0.8371	0.7559	0.7944
7	Maximizing recall	0.8559	0.7794	0.8159	0.6667	0.3789	0.4832	0.844	0.7406	0.7889
8	3BestRuns3BestTeams	0.8662	0.7613	0.8104	0.6792	0.3789	0.4865	0.8542	0.7242	0.7839
9	ExceptUColorado	0.8714	0.759	0.8114	0.6522	0.3158	0.4255	0.8333	0.7661	0.7983
10	ExceptUWM_UColorado	0.8439	0.8009	0.8218	0.6885	0.4421	0.5385	0.8333	0.7661	0.7983
11	ExceptUC3M_UColorado	0.8843	0.7523	0.813	0.7442	0.3368	0.4638	0.8767	0.712	0.7858
12	ExceptSCAI_UColorado	0.8521	0.7692	0.8086	0.7091	0.4105	0.52	0.8429	0.7344	0.7849
13	ExceptUCM_UColorado	0.8612	0.7579	0.8063	0.6981	0.3895	0.5	0.8508	0.7222	0.7812

We also performed some experiments using a union voting strategy. In this strategy, if a candidate pair is classified as DDI by at least one system, then the final decision for this pair will be DDI. A pair will be classified as negative only when it is classified as negative by all the systems. This strategy achieves a significant improvement of the recall (above 90%), but at expense of a strong decrease in precision (under 55%), and thereby, the *F*-measure also suffers a significant decrease (no more than 69%). Therefore, we decided not to use this strategy in our ensemble system.

As mentioned above, we conducted numerous experiments using different combinations of the final submissions. Table 22 shows the experimental results of some of these ensembles. For example, the second row shows the results obtained by combining all the runs. Since the UWM-TRIAD team did not provide any prediction for a total of 202 pairs, we decided to consider their unseen pairs as non-DDIs. In general, the ensemble systems do not overcome the FBK-irst system (see row 1 in Table 22). Therefore, we can conclude that the FBK-irst system is considerably more robust than the other systems.

We decided to evaluate whether removing the predictions of some teams had a positive or negative effect on the final performance. For example, we removed the predictions provided by the UCOLORADO_SOM team since its performance is markedly below that of the rest of the teams. However, this did not achieve an improvement in the results (see row 9 in Table 22 of the ensemble system, particularly on the MedLine dataset).

The ensemble made by the best run of each team (see row 6 in Table 22) achieves very close results to those reported by the best system, but not better than them. We also conducted an experiment in which we selected, for each team, the run that maximized recall (see row 7 in Table 22). This experiment showed lower results than the previous experiment.

Only one ensemble system (see the "6bestruns" row in Table 22) manages to improve the results of the FBK system very slightly. This ensemble consists of the 6 best submitted runs, that is, the runs submitted by the FBk-irst and WBI-DDI teams. This may indicate that if the FBK-irst system is extended to integrate the kernels proposed by the WBI teams, its performance may be improved. However, the difference between this ensemble and the FBK-irst system does not seem to be statistically significant.

8. Conclusion and future directions

The goal of DDIExtraction is to promote the development of information extraction techniques applied to the detection of drug names and DDIs from biomedical texts. There were a total of 38 runs which were submitted by 14 different teams from 7 different countries (6 of the teams participated in the drug name recognition task, while 8 participated in the DDI extraction task). The highest *F*1 scores obtained was 71.5% for drug name recognition and classification and 65.1% for extraction and classification of DDIs. This paper focuses on the extraction of DDIs. We have presented the main approaches used and examined the main challenges, which have yet to be resolved.

As regards the task of detection of DDIs, the participating systems demonstrated substantial progress over the previous DDI Extraction 2011. The best team, FBK-irst, achieved a competitive *F*-measure of 82.7% on DrugBank texts. However, performance on MedLine was lower mainly due to the limited size of its training dataset. Another possible reason may be that MedLine texts have a greater complexity than DrugBank texts. All teams used machine learning methods, specifically SVM. In general, non-linear kernel-based methods overcome linear SVMs.

We conclude that research into DDI extraction must continue. The error analysis points out the main limitations of the participating systems. Current approaches have focused on syntactic aspects, drawing their attention to the sentence structure. The resolution of linguistic phenomena such as cataphora, anaphora, appositive and coordinate structures and complex sentences, among others, could lead to better performance.

On the other hand, few participating systems took into account the sentence meaning. Approaches using domain knowledge have been recently applied with success to the pharmacological domain [57,58]. The use of knowledge resources can reduce the number of false positives generated by the current DDI extraction systems because these resources can help to distinguish between those pairs of drugs that are DDIs from those that are not. The information required for a semantic-based IE system can be taken, for example, from pharmacological databases such as DrugBank, PharmGKB [59], SIDER [60] or KEGG [61], among others. Some of them describe specific pairs of interacting drugs. For example, in DrugBank 39 different drugs that interact with *ciprofloxacin* are described. On the other hand, a larger number of DDIs can be deduced indirectly by exploiting, for example, the drug-protein relationships. Thus, the relationships of two different drugs with the same protein can be used to infer the mechanism leading to a DDI [62]. For example, *ciprofloxacin* is described to inhibit the activity of the metabolic enzyme *CYP1A2*, and *duloxetine* is described to be metabolized by *CYP1A2*. Therefore, there could be an interaction between *ciprofloxacin* and *duloxetine*. Similarly, the relationships of two different drugs with the same adverse drug reaction (ADR) can be used to infer possible DDIs [63]. For example, *morphine* is related to the side effect *central nervous system depression*. Therefore, other drugs related to the same ADR, such as *oxycodone*, could interact with *morphine*.

Up to now, the main limitation for the development of semantic-based approaches has been the availability of appropriate knowledge bases in a machine-readable format. However, the creation of these knowledge bases is becoming more feasible and common in the pharmacological domain [64,65]. This is due to the increasing number of databases and web servers providing structured and semi-structured pharmacological information, such as DrugBank or KEGG. Moreover, there are different community projects such as BIO2RDF [66] or LODD [67], which work to link the various sources of biological and pharmacological data together, enabling the integration of several pharmacological aspects described in different databases [68]. Another important factor is the proliferation of biomedical ontologies to store and formally represent domain knowledge. Ontologies enable the integration of the information disperse through different and heterogeneous databases and provide artifacts that can be exploited by IE systems [69].

Therefore, future directions for DDI extraction might entail the combination of syntactic and semantic information. In addition, increasing the size of training dataset, in particular for MedLine, would also have a very positive impact on the results.

Acknowledgments

Funding: This work was supported by the EU project TrendMiner [FP7-ICT287863], by the project MULTIMEDICA [TIN2010-20644-C03-01] and by the Research Network MA2VICMR [S2009/TIC-1542].

References

[1] W.H.O. (WHO), The importance of pharmacovigilance: safety monitoring of medicinal products.
[2] Bond C, Raehl CL. Adverse drug reactions in united states hospitals. Pharmacotherapy 2006;26(5):601–8.
[3] Leendertse AJ, Egberts AC, Stoker LJ, van den Bemt PM, et al. Frequency of and risk factors for preventable medication-related hospital admissions in the Netherlands. Arch Internal Med 2008;168(17):1890–6.
[4] Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. PLoS one 2009;4(2):e4439.
[5] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients. JAMA: J Am Med Assoc 1998;279(15):1200–5.
[6] Businaro R. Why we need an efficient and careful pharmacovigilance? J Pharmacovigilance 2013;1(4):1000e110.
[7] Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. BMJ: Br Med J 2004;329(7456):44.
[8] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;34(Suppl. 1):D668–72.
[9] Rodríguez-Terol A, Caraballo M, Palma D, Santos-Ramos B, Molina T, Desongles T, et al. Quality of interaction database management systems. Farmacia Hospitalaria 2009;33(3):134–46.
[10] Paczynski RP, Alexander GC, Chinchilli VM, Kruszewski SP. Quality of evidence in drug compendia supporting off-label use of typical and atypical antipsychotic medications. Int J Risk Saf Med 2012;24(3):137–46.
[11] Aronson J. Communicating information about drug interactions. Br J Clin Pharmacol 2007;63(6):637–9.
[12] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinform 2005;6(Suppl. 1):S1.
[13] Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 2008;9(Suppl. 2):S1.
[14] Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A. An overview of BioCreative II. 5. IEEE/ACM Trans Comput Biol Bioinform 2010;7(3):385–99.
[15] Arighi C, Lu Z, Krallinger M, Cohen K, Wilbur W, Valencia A, et al. Overview of the BioCreative III workshop. BMC Bioinform 2011;12(Suppl. 8):S1.
[16] Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Association for Computational Linguistics; 2009. p. 1–9.
[17] Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii J. Overview of BioNLP shared task 2011. In: Proceedings of the BioNLP shared task 2011 workshop. Association for Computational Linguistics; 2011. p. 1–6.
[18] Nédellec C, Bossy R, Kim J-D, Kim J-j, Ohta T, Pyysalo S, et al. Overview of BioNLP shared task 2013. ACL 2013;2013:1.
[19] Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. In: AMIA annual symposium proceedings. AMIA Symposium; American Medical Informatics Association; 2007. p. 1252–3.
[20] Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.
[21] Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Information access evaluation. Multilinguality, multimodality, and visualization. Springer; 2013. p. 212–31.
[22] Segura-Bedmar I, Martınez P, Sánchez-Cisneros D. The 1st DDIExtraction-2011 challenge task: extraction of drug–drug interactions from biomedical texts. Challenge Task Drug–Drug Interact Extr 2011;2011:1–9.
[23] Segura-Bedmar I, Martıinez P, Herrero-Zazo M. SemEval-2013 Task 9: extraction of drug–drug interactions from biomedical texts. In: Proceedings of the 7th international workshop on semantic evaluation (SemEval 2013); 2013.
[24] Stenetorp P, Topić G, Pyysalo S, Ohta T, Kim J-D, Tsujii J. Bionlp shared task 2011: supporting resources. In: Proceedings of BioNLP shared task 2011 workshop. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 112–20. <http://www.aclweb.org/anthology/W11-1816>.
[25] Cohen J et al. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
[26] Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform 2013;46(5):914–20.
[27] Chowdhury MFM, Lavelli A. Exploiting the scope of negations and heterogeneous features for relation extraction: a case study for drug–drug interaction extraction. In: Proceedings of NAACL-HLT. Curran Associates, Inc.; 2013. p. 765–71.
[28] Chowdhury MFM, Lavelli A. Impact of less skewed distributions on efficiency and effectiveness of biomedical relation extraction. In: COLING (Posters). Indian Institute of Technology Bombay; 2012. p. 205–16.
[29] Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. In: EACL, vol. 2006. The Association for Computer Linguistics; 2006. p. 98–113.
[30] Moschitti A. A study on convolution kernels for shallow semantic parsing. In: Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics; 2004. p. 335.
[31] Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T. All-paths graph kernel for protein–protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinform 2008;9(Suppl. 11):S2.
[32] Vishwanathan S, Smola AJ. Fast kernels for string and tree matching. Kernel Methods Comput Biol 2004:113–30.

[33] Moschitti A. Making tree kernels practical for natural language learning. In: EACL. The Association for Computer Linguistics; 2006.
[34] Kuboyama T, Hirata K, Kashima H, Aoki-Kinoshita KF, Yasuda H. A spectrum tree kernel. Inform Media Technol 2007;2(1):292–9.
[35] Björne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T. Extracting complex biological events with rich graph-based feature sets. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Association for Computational Linguistics; 2009. p. 10–8.
[36] Neves ML, Carazo JM, Pascual-Montano A. Extraction of biomedical events using case-based reasoning. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Association for Computational Linguistics; 2009. p. 68–76.
[37] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
[38] Fellbaum C. WordNet. Wiley Online Library; 1999.
[39] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77.
[40] Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: a library for large linear classification. J Mach Learn Res 2008;9:1871–4.
[41] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2(3):27.
[42] Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st annual meeting on association for computational linguistics, vol. 1. Association for Computational Linguistics; 2003. p. 423–30.
[43] Charniak E, Johnson M. Coarse-to-fine n-best parsing and Maxent discriminative reranking. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics; 2005. p. 173–80.
[44] McClosky D, Adviser-Charniak E. Any domain parsing: automatic domain adaptation for natural language parsing. Ph.D. thesis; 2010.
[45] Chowdhury MFM, Lavelli A. Disease mention recognition with specific features. In: Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics; 2010. p. 83–90.
[46] Liu H, Christiansen T, Baumgartner Jr WA, Verspoor K. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. J Biomed Semantics 2012;3:3.
[47] Coulet A, Garten Y, Dumontier M, Altman RB, Musen MA, Shah NH, et al. Integration and publication of heterogeneous text-mined relationships on the semantic web. J Biomed Semantics 2011;Suppl. 2(2):S10.
[48] Schmid H. Improvements in part-of-speech tagging with an application to German. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, editors. Natural language processing using very large corpora. Dordrecht, Netherlands: Kluwer Academic Publishers.; 1999. p. 13–26.
[49] Paice CD. Another stemmer. ACM SIGIR Forum 1990;24(3):56–61.
[50] Porter MF. An algorithm for suffix stripping. Program: Electron Libr Inform Syst 1980;14(3):130–7.
[51] Lease M, Charniak E. Parsing biomedical literature. In: Natural language processing – IJCNLP 2005. Springer; 2005. p. 58–69.
[52] Sagae K, Tsujii J. Dependency parsing and domain adaptation with LR models and parser ensembles. In: EMNLP-CoNLL. The Association for Computational Linguistics; 2007. p. 1044–50.
[53] Hunter L, Lu Z, Firby J, Baumgartner WA, Johnson HL, Ogren PV, et al. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC Bioinform 2008;9(1):78.
[54] Zhou X, Zhang X, Hu X. Dragon toolkit: incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. 19th IEEE international conference on tools with artificial intelligence. ICTAI 2007, vol. 2. IEEE; 2007. p. 197–201.
[55] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–7.
[56] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895–923.
[57] Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics 2010;11(10):1467–89.
[58] Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinform 2014;15(1):64.
[59] Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res 2002;30(1):163–5.
[60] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6(1).
[61] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40(D1):D109–14.
[62] Hage DS, Tweed SA. Recent advances in chromatographic and electrophoretic methods for the study of drug-protein interactions. J Chromatogr B: Biomed Sci Appl 1997;699(1):499–525.
[63] Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321(5886):263–6.
[64] Khelashvili G, Dorff K, Shan J, Camacho-Artacho M, Skrabanek L, Vroling B, et al. GPCR-OKB: the G protein coupled receptor oligomer knowledge base. Bioinformatics 2010;26(14):1804–5.
[65] Whirl-Carrillo M, McDonagh E, Hebert J, Gong L, Sangkuhl K, Thorn C, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 2012;92(4):414–7.
[66] Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008;41(5):706–16.
[67] Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminform 2011;3(1):19.
[68] Pathak J, Kiefer RC, Chute CG. Using linked data for mining drug–drug interactions in electronic health records. Stud Health Technol Inform 2013;192:682.
[69] Wimalasuriya DC, Dou D. Ontology-based information extraction: an introduction and a survey of current approaches. J Inform Sci 2010;36(3):306–23.